DOCUMENT RESUME

ED 049 290 TM 000 477

AUTHOR Crocker, Linda M.; Mehrens, William A.
TITLE The Comparative Effectiveness of Different Item Analysis Techniques in Increasing Change Score Reliability.
PUB DATE Feb 71
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Attitude Tests, Beliefs, *Changing Attitudes, Comparative Analysis, *Individual Development, *Item Analysis, Longitudinal Studies, Research Tools, *Scores, *Student Attitudes, Test Reliability
IDENTIFIERS *Inventory of Beliefs

ABSTRACT

Four new methods of item analysis were used to select subsets of items which would yield measures of attitude change. The sample consisted of 263 students at Michigan State University who were tested on the Inventory of Beliefs as freshmen and retested on the same instrument as juniors. Item change scores and total change scores were computed for each subject. Responses of half the sample were used for item analyses. The four methods of change item analysis employed were: selection on the basis of high change item score variance; selection on the basis of pretest response frequency; selection on Saupe's correlation between change item score and total change score; and selection on triserial correlation between item change score and total change score. Subsets of 15, 30, 60, and 90 items were chosen by each method. In addition, subsets of equal size were randomly selected. When change score reliability was computed, using the responses of the cross-validation group, all four methods of item analysis resulted in higher change score reliability than did random selection. (Author)

THE COMPARATIVE EFFECTIVENESS OF DIFFERENT ITEM ANALYSIS TECHNIQUES IN INCREASING CHANGE SCORE RELIABILITY

Linda M. Crocker

University of Florida

William A. Mehrens

Michigan State University

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 049290

TM 000 427

A methodological problem frequently encountered by researchers in education is how to obtain measures of growth or change for individuals over a given period of time. One approach to this problem has been to calculate the change score for each individual, using the formula:

$$D = Y - X$$

where D is the change score, Y is the score at time 2, and X is the score at time 1. Since both X and Y are totals of individual item scores on each occasion, it is possible to define a change item score as

$$d_i = y_i - x_i$$

where d_i is an individual's change score on a single item, x_i is the individual's score on item i at time 1 and y_i is his score on that item at time 2. Thus total change score may be defined as

$$D = \Sigma d_i = \Sigma(y_i - x_i).$$

Researchers who have attempted to use such change scores have been plagued by one persistent psychometric problem. These change scores are remarkably unreliable (Harris, 1963). When the researcher is primarily interested in measuring change for a group, this problem of low reliability is not too serious; however, if he wishes to make meaningful comparisons between individuals on the basis of their growth or attitude change, then the lack of reliability becomes crucial.

Ordinary item analysis procedures, usually based upon a single test administration, are designed to improve test internal consistency or to yield a test which correlates highly with some criterion. Such methods are not guaranteed to enhance change score reliability. Theorists such as Bereiter (1963), Saupe, (1961 and 1966), and Lord (1968, p. 331) have suggested that a researcher who wishes to construct an instrument, sensitive to individual change, should use item analysis techniques expressly suited for that purpose.

Several new techniques for change item analyses have been developed recently. Among these are: selection of items with high change item score variance, $S_{d_i}^2$; selection on the basis of pretest response frequency; selection on the basis of of high r_{d_D} values when item change score is correlated with total change score; and selection of items with high values of r_{tris}, when the triserial correlation between item change score and total change score is computed. (The latter method is only applicable when items are dichotomously scored at time 1 and time 2 so that change item scores can only assume values of -1. 0, or 1.)

Because of the relative newness of these item analysis methods there has been little empirical research to demonstrate whether or not they could successfully increase change score reliability. Also the comparative efficiency and effectiveness of these different procedures has been virtually unknown.

Purpose of this Study.

The purpose of this study was to determine whether the use of the item analysis methods previously discussed would increase the reliability of change scores on a collegiate attitude survey. The investigation was designed to answer three specific questions:

1. Could subsets of items, chosen by any of these methods of change item analysis, have higher change score reliability than subsets of items chosen by random selection?
2. How could these methods be ranked in terms of their ability to improve change score reliability?
3. Could reliable sets of change items be selected on the basis of pretest data alone?

This last question was considered singularly important because of its practical significance for test construction. In many attempts to measure change the experimenter simply does not have time to construct his instrument and run a complete item analysis on test-retest data before proceeding with the experiment. (This is especially true with longitudinal studies.) Thus, if a method could be developed to eliminate less useful items on the basis of pretest characteristics alone, it would be extremely helpful and time-saving for the researcher and his subjects.

Methodology.

The sample used for this study was a group of 263 students at Michigan State University who were tested on a battery of aptitude, attitude, and interest measures in their freshman year, and who were retested on these measures as juniors three years later. The instrument employed in this study was the Inventory of Beliefs, Form I. This attitude survey was developed by the Cooperative Study of Evaluation in General Education under the sponsorship of the American Council on Education Committee on Measurement and Evaluation. The scale was designed to measure an individual's tendency to endorse stereo-typic beliefs (Lehmann and Dressel, 1963). Some sample items from this inventory are:

"No world organization should have the right to tell Americans what they can or cannot do."

"We would be better off if there were fewer psychoanalysts probing and delving into the human mind."

" Parents know as much about how to teach children as public school teachers."

There were four possible responses to each item--Strongly Agree, Agree, Disagree, and Strongly Disagree. Two separate scoring schemes were used. Under the first system the examinee was awarded one point for each Disagree or Strongly Disagree response; with the second method, a one-to-four scoring scheme was used, ranging from one point for Strongly Agree to four points for Strongly Disagree.

The experiment was conducted with an item analysis, cross-validation design. The sample was randomly split into two groups with 132 students assigned to the item analysis group and 131 students assigned to the cross validation group.

Item Analysis Procedures.

Method I was an item analysis procedure requiring the selection of items with high change score variance, $S_{d_i}^2$. This tends to eliminate those items for which the group exhibited little change over time as well as items for which there was a universal response shift, retaining items on which there was a high degree of individual variability between subjects in response change. After the change item scores, d_i, were computed, the mean change score, $\bar{d}_i$, and the change score variance $S_{d_i}^2$ were computed for each item. Items with the largest values of $S_{d_i}^2$ were selected. Subsets of 15, 30, 60, and 90 items were chosen from the original pool of 120 items.

Method II required that items be chosen on the basis of pretest response frequency. With this method it was necessary to take into account the expected

direction of the change. Suggested by Gruber and Weitman (1962), this method required the selection of items which had a low percentage of positive responses on the pretest, if is was known that a high percentage of positive responses could be expected on the post test, or vice versa. Because the Inventory of Beliefs had been developed to measure attainment of objectives of higher education, it seemed reasonable to predict that students' scores would increase over time. (Data from Lehmann's and Dressel's study upheld this prediction.) Item means, $\bar{x}_i$, were computed for each item on the pretest. Items with the lowest mean scores were selected into the 15, 30, 60, and 90 item subsets.

Method III was a correlational item analysis procedure for which the index of item selection was the expression derived by Saupe (1966):

$$r_{dD} = \frac{C_{xX} + C_{yY} - C_{xY} - C_{Xy}}{\sqrt{S_x^2 + S_y^2 - 2C_{xy}}\ \sqrt{S_X^2 + S_Y^2 - 2C_{XY}}}$$

where x and y denote item scores, X and Y are total scores and C is covariance. Items which had the greatest positive correlations with total change score were selected into the test subsets.

For Method IV the triserial correlation coefficients (Jenkins, 1956) between the trichotomized item change score and total change score were computed according to the formula:

$$r_{tris} = \frac{M_1 y_1 + M_0(y_{-1} - y_1) - M_{-1} y_{-1}}{\sigma \left[\frac{y_1^2}{p_1} + \frac{(y_{-1} - y_1)^2}{p_0} + \frac{y_{-1}^2}{p_{-1}} \right]}$$

where M is mean total score, y is curve ordinate, σ is S.D. of the scores, and P is the proportion of examinees who had change item scores of -1, 0, or 1.

This method, of course, was only applied to the data that had been scored on a zero-one basis on the original tests. Items with the highest positive values for r_{tris} were selected into the test subsets.

The control method consisted of selecting randomly subsets of 15, 30, 60, and 90 items for comparison with those which had been chosen by the systematic item analysis procedures.

After the subsets of items had been selected, using data from the 132 students in the item analysis group, the change score reliability for each item subset was computed using the item responses of the 131 students in the cross validation group. The computational formula used to obtain the change score reliability estimates was a change-score version for computing coefficient alpha, or Kuder-Richardson 20, derived by Webster and Bereiter (1963):

$$r_{DD} = \frac{k}{k-1}\left[1 - \frac{\sum S^2 d_i}{\sum S_{d_i}{}^2 + \sum_{i \neq j} C_{d_i d_j}}\right]$$

where $S_{d_i}{}^2$ is change item variance, C is covariance, and k is the number of items.

RESULTS.

The change score reliability estimates computed for subsets of 15, 30, 60, and 90 items selected by each of the item analysis methods are presented in Table 1. From these results it is apparent that Saupe's method of change item analysis consistently resulted in more reliable subsets of items than did either of the other two item analysis methods or the control method of random selection when items were scored on the one-to four point scale.

Table 1. Change score reliability coefficients computed for the cross validation sample using the one-to-four scoring system.

Item Analysis Method	Number of Items			
	15	30	60	90
Method I (Change Variance)	.50	.61	.75	.83
Method II (Pretest Frequency)	.50	.65	.78	.83
Method III (Saupe's r_{dD})	.63	.70	.80	.85
Random Selection	.30	.49	.70	.80

There was little difference between the reliability coefficients of item subsets chosen by the two response frequency methods (Method I and Method II); however, both of these methods resulted in higher reliability of change scores than did the control method for subsets of 15, 30, 60, and 90 items. Another point that should be noted from the data presented in Table 1 is that the differences between reliability coefficients were greater when fewer items were selected from the original pool. At the 90-item level the reliability values ranged from only .85 to .80. At the 15-item level, however, the range was from .63 to .30.

When the items on the attitude survey were scored on a zero-one basis, it was possible to introduce a fifth method of item selection (triserial correlation) in addition to the three item analysis methods used for one-to-four scoring and random selection. Change score reliabilities for the 15, 30, 60, and 90 item subsets were computed as before, using the responses of the cross validation group. These change score reliability estimates are presented in Table 2.

Table 2. Change score reliability coefficients computed for the cross validation sample using the zero-one scoring system.

Item Analysis Method	Number of Items			
	15	30	60	90
Method I (Change Variance)	.52	.56	.68	.72
Method II (Pretest Frequency)	.36	.52	.67	.72
Method III (Saupe's r_{dD})	.33	.49	.68	.74
Method IV (triserial r)	.37	.56	.68	.75
Random Selection	.21	.48	.57	.67

The differences between the methods of item analysis were much less pronounced under this scoring system. In general, however, all four methods of change item analysis consistently resulted in higher estimates of change score reliability than did the technique of random selection. The greatest differences again were observed when fewer items were selected from the original pool.

CONCLUSIONS AND PRACTICAL IMPLICATIONS.

From a practitioner's viewpoint, several of the findings of this study can be applied to the area of constructing instruments to measure change. First, it appears that change item analysis can be a profitable approach to solving the change score reliability problem. Certainly researchers should consider using these techniques when constructing new instruments to measure change or when forced to shorten an already existing scale.

Second, the use of a multiple-response format for items (e.g. a one-to-four point scoring system) seems to allow a more sensitive observation of change and makes the selection of a method of change item analysis an important consideration.

A third point, having great significance for the longitudinal researcher, is that considerable time and expense might be saved by selecting items on the basis of pretest response alone. It should be remembered, however, that this can only be done when the direction of change can be predicted in advance.

References:

1) Bereiter, Carl M. Some persisting dilemmas in the measurement of change. Chapter I in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

2) Gruber, H. E., and Weitman, M. Item analysis and the measurement of change. Journal of Educational Research, 1962, 6, 287-289.

3) Jenkins, William L. Triserial r--a neglected statistic. Journal of Applied Psychology, 1956, 40, 63-64.

4) Lehmann, Irvin J., and Dressel, Paul L. Changes in critical thinking, attitudes, and values associated with college attendance. Final Report of Cooperative Research Project No. 1646. East Lansing: Michigan State University, 1963.

5) Lord, Frederick M. Elementary models for measuring change. Chapter 2 in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

6) Saupe, Joe L. Technical considerations in measurement. Appendix in Dressel, Paul L. (Ed.) Evaluation in higher education. Boston, Mass.: Houghton Mifflin, 1961.

7) Saupe, Joe L. Selecting items to measure change. Journal of Educational Measurement, 1966, 3, 223-228.

8) Webster, Harold, and Bereiter, Carl. The reliability of changes measured by mental test scores. Chapter 3 in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.